From passive to interactive object learning and recognition through self-identification on a humanoid robot

Natalia Lyubova, Serena Ivaldi, David Filliat

Noname manuscript No.
(will be inserted by the editor)

From passive to interactive object learning and recognition through self-identification on a humanoid robot

N. Lyubova, S. Ivaldi,
D. Filliat

Received: date / Accepted: date

Abstract Service robots, working in evolving human environments, need the ability to continuously learn to recognize new objects. Ideally, they should act as humans do, by observing their environment and interacting with objects, without specific supervision. Taking inspiration from infant development, we propose a developmental approach that enables a robot to progressively learn objects appearances in a social environment: first, only through observation, then through active object manipulation. We focus on incremental, continuous, and unsupervised learning that does not require prior knowledge about the environment or the robot. In the first phase, we analyse the visual space and detect proto-objects as units of attention that are learned and recognized as possible physical entities. The appearance of each entity is represented as a multi-view model based on complementary visual features. In the second phase, entities are classified into three categories: parts of the body of the robot, parts of a human partner, and manipulable objects. The categorization approach is based on mutual information between the visual and proprioceptive data, and on motion behaviour of entities. The ability to categorize entities is then used during interactive object exploration to improve the previously acquired objects models. The proposed system is implemented and evaluated with an iCub and a Meka robot learning 20 objects. The system is able to recognize objects with 88.5% success and create coherent representation models that are further improved by interactive learning.

Keywords Developmental robotics · interactive object learning · self-identification · object recognition

N. Lyubova
Aldebaran-Robotics - Perception team, Paris, France.
E-mail: nlyubova@aldebaran.com

S. Ivaldi
Inria, Villers-lès-Nancy, F-54600, France &
IAS, TU-Darmstadt, Germany &
Institut des Systèmes Intelligents et de Robotique, CNRS UMR 7222 & Université Pierre et Marie Curie, F-75005, France.
E-mail: serena.ivaldi@inria.fr

D. Filliat
ENSTA ParisTech - INRIA FLOWERS Team, Computer Science and System Engineering Laboratory, ENSTA ParisTech, Paris, France.
E-mail: david.filliat@ensta-paristech.fr

1 INTRODUCTION

Robots are coming into everyday life, not only as factory robots but also as service robots helping people to increase the performance of their work and improve the quality of their life. While factory robots work in well-structured environments, service robots or personal robots will work in human environments that are less predictable and less structured. These robots will need the ability to adapt to changing situations and continuously learn new information about the surrounding environment. Moreover, these robots should be able to learn without constant human supervision, and learning should be autonomous and continuous, with the possibility of using discontinuous interactions with humans and performing autonomous actions in order to acquire information. Among many different skills, a robot working in a human environment should be able to perceive the space around it in order to identify meaningful elements such as parts of its own body, objects, and humans. In this paper, we focus on the issue of learning the appearances and recognizing the elements that appear in the working space of a humanoid robot, and we call these elements *physical entities*. Learning is performed based on both passive observation when a human manipulates objects in front of the robot and interactive actions of the robot (Fig. 1).

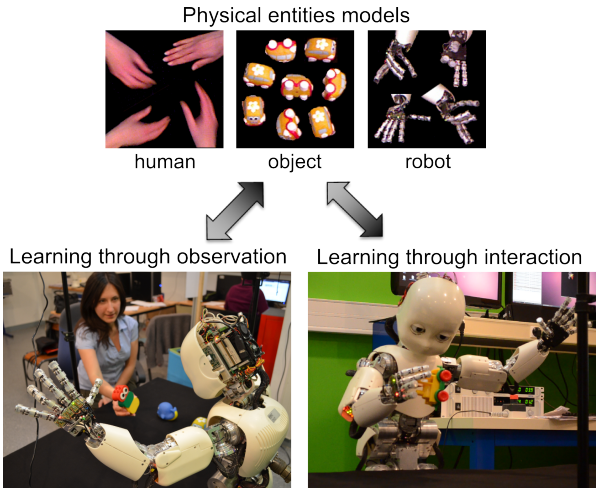


Fig. 1 The main modules of the proposed approach: learning through passive observation when a human manipulates an object in front of the robot and learning through interactive actions of the robot.

Various computer vision approaches achieve good performances for detecting specific physical entities of particular classes, like human faces [69], skin parts [73], coloured [20] or textured [3] objects. Most of these approaches are based on prior knowledge, either assuming very specific objects (such as human hands [72] or robot hands [47] of particular color, or by using artificial markers [17]) or requiring carefully created image databases, where images of each object are labeled in order to perform supervised learning. For example, the organizers of the Pascal VOC challenge [14] put a lot of effort in creating and improving image databases that were very beneficial to algorithms performance over years. Other approaches include a specific object learning phase, for example using a turntable to rotate an object and learn its appearance from different viewing angles. Prior knowledge and supervision facilitate object detection, but they are not easily applicable for autonomous robots that need to adapt to different human users and new objects at any time. Indeed, in such setup, specific or supervised approaches limit the adaptability of the robot, since it is difficult to extend these approaches for online continuous detection and learning of new objects without specific human supervision. Therefore, we propose that object recognition in this context should be based on general high-level representations and learning methods that could be applied to all physical entities of the environment and could support learning by observation and by interaction.

The human development is a very motivating example of efficient learning about the environment without explicit supervision. Indeed, object representation is considered as one of the few core knowledge that form the basis of human cognition [66]. It is interesting to note that these capabilities are acquired progressively through a long period during infancy that plays an important role in human life. At first, a baby learns mostly through observation, because of its limited manipulation capabilities, in an environment where the parents are present most of the time. Thus the social environment is the cause of a large part of the sensory stimulus, even if the social engagement of the baby remains limited. Progressively, the baby learns about his own body, and its control, which then makes it possible to manipulate objects [53]. It has been shown in many studies (e.g. [26]) that such capability improves knowledge of the surrounding world and in particular the objects. The social interactions then take a growing importance as learning focus on more complex activities. Infant development has inspired a variety of research studies on autonomous robots learning. The characteristics of infants learning process, such as being continuous, incremental, and multi-modal are reflected in different approaches in developmental robotics [71]. In contrast to traditional robotics, a developmental approach does not focus on a fast achievement of predefined goals, but rather on an open-ended learning process, where the performance improves over time, the learning process being flexible and allowing to adapt to changing circumstances.

In this paper, we propose a developmental approach taking inspiration from the human development related to object appearance learning and recognition [65]. We describe a perceptual system that makes it possible for a robot to learn about physical entities in its environment in a two stage developmental scenario (Fig. 1):

1. *learning by observation*: the robot learns appearance models of moving elements, where the motion is mostly produced by a human partner who demonstrates different objects,
2. *interactive learning*: the robot interacts with objects in order to improve its knowledge about objects appearances after having identified the parts of its own body, parts of a human partner, and manipulable objects.

Our main contribution consists in the integration of a generic perception capability, self- and others- identification, and interactive actions for active exploration of the surrounding environment and its objects. Our algorithm requires very limited prior knowledge and does not require predefined objects, image databases for learning or dedicated detectors, such as markers or human face/skin/skeleton detectors. Instead, using a color and depth camera (hereafter called RGB-D sensor), the visual space is autonomously segmented into physical entities whose appearances are continuously and incrementally learned over time and synthesized into multi-view representation models. All entities are then categorized into parts of the body of the robot, human parts, and manipulable objects, which make it possible to correctly update objects models during their manipulation. Note that even if the social interactions may take a large part in the learning of objects, it is not the subject of the current paper, but we refer to our previously published work on socially-guided learning, where the robot learns objects with a human partner providing additional feedback used to guide learning [32], [51].

The paper is organized as follows: Section II gives a brief overview of related work on unsupervised learning and interactive learning including self- and others- identification; the proposed perceptual approach of learning through observation is detailed in Section III and the interactive learning approach is described in Section IV; the experimental evaluation is reported in Section V; and Section VI is devoted to discussion of the results.

2 Related work

We are working on unsupervised object learning and interactive perception as a generic approach towards autonomous learning integrating perception and control. Object learning has been addressed in a huge number of computer vision

approaches whose exhaustive review is outside the scope of this paper (see [23] or [14], for example). We will therefore restrict ourselves to the approaches closely related to our algorithmic choices. Interactive perception has been used for detecting and segmenting objects in a scene, for learning objects properties and appearances, or exploring affordances. Moreover, some studies on interactive perception integrate identification of parts of the robot (especially hands) and use their localization to improve object segmentation or learning algorithms. We will not cover the more general area of learning by demonstration as our approach depends only on the entities motions produced by humans manipulating objects and used to learn appearances models, but does not rely on a detailed analysis of the human demonstrations and does not try to imitate the human behaviour.

2.1 Unsupervised object learning

In our approach, as suggested by studies on the development of object perception capabilities in humans [65], the perception of the environment begins by detection of meaningful elements in the visual field of the robot. These elements are detected from generic principles such as cohesion and continuity, while most traditional object detection approaches are based on prior knowledge or dedicated algorithms providing robust detection of specific objects of particular categories. More generic approaches segment a scene into coherent image regions and further segment objects from the background based on consistency of visual characteristics [64] or motion behaviour [54]. Similar principles have also been used to detect and model objects using laser range finders [46]. Other unsupervised approaches in vision are aimed at detecting not a concrete object, but an evidence of an object existence or a proto-object [55], [56]. Taking inspiration from human vision, a proto-object is defined as a unit of attention or a localized visual area with certain properties, representing a possible object or its part. Proto-object detection is often based on biologically motivated mechanisms of selective attention, for example visual saliency [52], [70].

Once an object or a proto-object is detected, its visual appearance is analyzed and often encoded within more compact descriptors characterizing local features or general visual content, like color or texture [6]. While balancing between robustness, speed, and the ability to preserve information, a good descriptor should allow to discriminate different objects and accommodate intra-object variations. Based on extracted features, an efficient object representation should characterize a significant part of the visual content in a short description. In order to improve recognition, object representation can combine several types of visual features. In this case, the efficiency of object recognition will be higher with complementary descriptors characterizing different types of visual data while avoiding redundancy [12].

A widely-used object representation methods is the Bag of Words (BoW). It represents objects or images as collections of unordered features quantized into dictionaries of visual words, and each object is encoded by its visual words. In this case, the learning procedure consists in training a classifier on extracted visual words, and the recognition procedure consists in applying the classifier on extracted visual words [63]. Among existing studies, there are many variations of BoW based on a pixel-level description [1], image patches [61], or local features for example, keypoints [63], [18], edges [16], and regions [57]. Instead of using a simple list of visual words, the importance of each visual word can be taken into account by using Term Frequency-Inverse document frequency (TF-IDF) approach [63]. In this case, an object is encoded by occurrence frequencies of its visual words, and TF-IDF approach is used to evaluate the importance of words with respect to objects and give higher weights to distinctive visual words. An inverted index allows to quickly compare each set of extracted visual words with all memorized objects.

The main weakness of BoW approaches is the absence of spatial relations between visual words inside images. This limitation is resolved in variations of BoW, like part-based models such as the Constellation model [15], or the k-fans model [10]. Part-based models combine appearance-based and geometrical models, where each part represents local visual properties, and the spatial configuration between parts is characterized by a statistical model or springconnections representing "deformable" relation between parts. These models are based on learning the geometrical relations between image parts or features, like local features [15] or edges [16].

2.2 Interactive learning

In the context of learning about the surrounding environment, some knowledge can be acquired through simple observation, without performing any action, through the image processing techniques reviewed in the previous section. However, it is not easy to bind all gathered information into coherent objects representations and learn the overall appearances of the objects. Actions of the robot provide an ability to detect manipulable objects in a scene, segment them from the background, better learn their overall appearances and properties, thus allowing to find out an appropriate way of interaction with these objects. Interactive actions are useful for both object learning and also object recognition in ambiguous situations, when dealing with several similar objects and when more evidences are needed for object identification [5].

Several approaches have been proposed to take advantage of interactive actions and various perceptual channels. For example, in [67], an unknown object is manipulated and tapped with the robot finger in order to produce a sound that

is used to recognize this object. The authors of [62] propose a more complex approach that integrates auditory and proprioceptive feedback when performing five different actions on a set of 50 objects, showing very high recognition rates. In [9], an advanced tactile sensor is used with five different exploratory procedures in order to associate haptic adjectives (i.e. categories like hard, soft,...) to objects. And in [24], the evolution of the visual motion of objects during robot actions are analysed to classify objects into two categories as a container/non container. All these approaches take advantage of the behaviour of the object during or after manipulations, and therefore they are not applicable in the scenario based on observation that we use as a first stage in this paper. They however could be used as an interesting complement with our system for integration of multi-modal information whenever the visual information is not sufficient for recognition.

We therefore focus on studies of interactive approaches aimed at learning visual objects appearances. In [49], an object model is learned when the robot approaches the object closer to the visual sensor and captures images at four positions and orientations of the object. In [68], an object representation is generated from snapshots captured from several viewpoints, while the object is intentionally placed by the robot to the center of its visual field, rotated, and segmented from the background using the pre-learned background model. In [5], an object representation is learned as a collection of its views captured at orientations that are selected to maximize new information about the object. The object segmentation consists of cropping a central part of a captured image and subtracting the pre-learned background. As a common limitation of these approaches, a robot does not detect or grasp an object by itself, but the object is provided by a human partner placing the object directly in the hand of the robot. This scenario simplifies the system, since it does not require object detection, localization, or a grasp planning.

Perception and action can be also integrated into autonomous object exploration performed without human assistance. In [11], a pushing behaviour is used to move objects lying on a table in order to improve visual object segmentation and observe different views. The resulting images are used to train a classifier using a Bag of Words representation. In [25], two simple actions primitives are used to spread piled lego blocks in order to be able to sort them. A more advanced scheme is proposed in [28] to decide which pushing to perform in order to segment cluttered scenes on a table using a complex probabilistic model. However, these two last approaches do not integrate object learning and recognition. In [36], a sophisticated vision system provides a set of 2D and 3D features that makes it possible to generate object grasping hypothesis. The successful grasping then allows to achieve precise object motion used to integrate features from several views in order to produce coherent 3D models. In [37], object manipulation is used to generate autonomously complete 3D models of objects using a RGB-D camera. An initial grasp is performed through heuristics, before moving the object following an algorithm optimizing the information gained by the new view. This approach relies mainly on 3D model matching using the dense data provided by the camera. In contrast, our interactive learning approach is not designed to improve object segmentation (and thus is limited in its capacity to segment clustered scenes), but to improve the object appearance models in order to provide additional representative views. Moreover, we do not seek to produce precise 3D models of objects, but rather use multiple view appearance models for their adaptability in presence of changing observation condition and capacity to represent deformable objects (which is however not tested in the current paper).

Most interactive object exploration approaches make use of knowledge about the body of the robot. This knowledge can concern the body structure for control and correspond to the concept of *body schema*, or the appearance of the body and correspond to the *body image* [27]. In [49], the hand tracking is used for fixation on the object during manipulation, whereas in [43], the hand localization is used to improve object segmentation. In [37], the precise 3D model of the robot hand is used to precisely localize objects and remove the robots parts from the objects models. Therefore, in interactive scenarios, the self-identification and localization of the parts of the robot in the visual field allow more efficient processing of visual information during and after interaction with objects. In our approach, we assume a very limited prior knowledge of the body of the robot and we show that, as far as perceptual learning is concerned, the raw motor values are sufficient to learn and continuously adapt a body image that is sufficient to learn about objects during manipulation.

2.3 Self- and others- discrimination

As explained before, knowledge about the body image of the robot provides advantages for interactive exploration of the environment. As an inspiration, child development and especially sensorimotor developmental stages demonstrate the importance of own body exploration. An infant starts to learn about the world from developing a sense of his own body, and later on performs interactive actions directed to exploration of the environment [53].

2.3.1 Robot self-discovery

Among the variety of studies on self-discovery for robots learning its body image, most of them are based on prior knowledge or resort to local approaches. Some strategies exploit a predefined motion pattern of the robot, a predefined appearance of the body, or a known body schema, such as the joint-link structure. For example, in [30], the hand of the robot is detected based on a grasped object of a known appearance, and the hand tracking is based on tracking the object. In [47], the identification of the hand of the robot is based on wearing a glove of known color. These techniques simplify the robot self-identification but impose some limitations. Since these algorithms are dependent on a fixed appearance or behaviour, they cannot be easily adapted to changes in the appearance or motion pattern of the robot. The independence on prior knowledge would enable to overcome these limitations and generalize the self-identification over new appearances and new end-effectors, like grasped tools.

In early studies, the detection of the hand of the robot was based on its motion [42]. The important limitation of this approach is an assumption of a single source of motion. However, in real environments, visual motion can be produced not only by the robot itself but also by other agents that can be robots or humans.

Considering visual motion as a response of an action, the visual motion that follows almost immediately after an action of the robot can be used as a cue to localize the parts of the robot in the visual field. Based on this principle, self-identification based on the time-correlation between an executed action and visual motion is performed in [43], [44], and [21]. In [44], localization of the hand is based on a pre-learned time delay between the initiation of an action and the emergence of the hand in the visual field. Assuming a single source of motion at a time, the hand is identified as a moving region appearing first within the pre-learned time window after the initiation of the action. In [43], localization of the hand is based on the amount of correlation between the velocity of the movement and the optical flow in the visual field. This method allows to identify the hand among multiple sources of motion without requiring a priori information about the hand appearance.

A developmental approach of identification of the body of the robot based on visuomotor correlation is proposed in [58]. Visuomotor correlation is estimated from proprioceptive and visual data acquired during head-arm movements. In the learning stage, the robot performs motor babbling and gathers the visual and proprioceptive feedback in terms of visual motion and changes of motors states. In case of high correlation, the moving region is identified as a part of the robot, and the visuomotor information, such as the body posture and visual features, is stored in the visuomotor memory. This self-identification method is also adaptable to extended body parts.

2.3.2 Identification of self and others

A generic method aimed at understanding a dynamic environment based on contingency is proposed in [21]. The method allows to discriminate actions performed by the robot from actions performed by other physical actors considering the time delays between the actions and the responses and their respective durations. Autonomous identification of the hand of the robot during natural interaction with a human is proposed in [35]. The approach is based on mutual information estimated between the visual data and proprioceptive sensing. The value of mutual information is used to identify which visual features in a scene are influenced by actions of the robot. Since the system is aimed at detecting parts of humans and robots, it is mainly focused on visual regions that are close to the visual sensor and regions moving with a high speed.

We are interested in the identification of the parts of the robot during natural human-robot interaction and also in the identification of the parts of human partners and possible objects. We therefore propose a generic identification algorithm that is independent on the appearance and motion pattern of the robot and capable of identifying these three categories. This algorithm will be integrated with interactive object exploration in order to enhance the learning process.

3 Learning by observation

In this section, we describe the first stage of the proposed developmental approach that allows a robot to detect physical entities in its close environment and learn their appearances during demonstration by a human partner. Our approach is based on online incremental learning, and it does not require image databases or specialized face/skin/skeleton detectors. All knowledge is iteratively acquired by analyzing the visual data. Starting from extraction of low-level image features, the gathered information is synthesized into higher-level representation models of physical entities. Given the localization of the visual sensor, the visual field of the robot covers the interaction area including parts of the body of the robot, parts of a human partner, and manipulated objects.

In this work, we have chosen to use a Kinect RGB-D sensor (Kinect, [74]) instead of using stereo-vision based on the cameras in the robot eyes. Our choice is justified by the efficiency and precision of the RGB-D data, since the Kinect sensor allows fast acquisition of reasonably accurate depth data as will be discussed in Section 6. Both RGB and depth data are acquired with the OpenNI library[1]. Depth data are

[1] http //openni.org

only used during proto-object detection procedure to refine boundaries of possible objects. The overall algorithm of object learning can therefore work without the optional step requiring depth data, but rather based on RGB data only that could be performed with the embedded visual sensor.

3.1 Segmentation of the visual space into proto-objects

Learning about the close environment of the robot begins by segmenting the visual space into proto-objects as salient units of attention that correspond to possible isolated or connected physical entities. The main processing steps towards detection and segmentation of proto-objects are shown in Fig. 2.

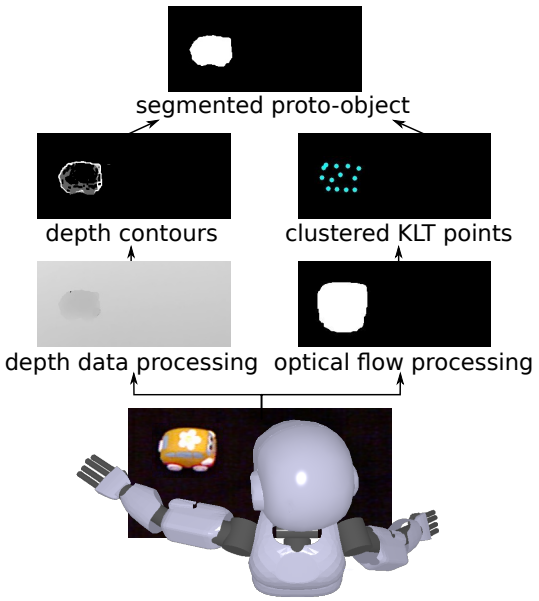


Fig. 2 Detection and segmentation of proto-objects.

Our proto-object detection approach relies on motion-based visual attention, since motion carries a significant part of information about events happening in the environment and their actors [22]. In our scenario, moving regions in the visual field correspond mainly to parts of the body of the robot, parts of a human partner, and manipulated objects, which are the entities we seek to learn. Moreover, in the case of motion-based visual attention, a human partner can attract the attention of the robot by simply interacting with an object in order to produce visual motion.

Our motion detection algorithm is based on the Running average[2] and image differencing. After detecting moving pixels, we fill holes and remove noisy pixels by applying the erosion and dilation operators from mathematical morphology [60]. Further, based on the constraints of the working area of the robot, we ignore the visual areas that are unreachable for the robot.

The detected moving regions of the visual field are analyzed as probable locations of proto-objects. Inside each moving region, we extract Good Features to Track (GFT) [59] developed especially for a tracking purpose. The extracted GFT points are tracked between consecutive images using the Lucas-Kanade method [40] chosen due to its small processing cost, accuracy, and robustness. We analyse the motion behaviour of tracked points in order to detect areas of uniform motion, which allow to isolate proto-objects inside moving image regions. Tracked points are grouped into clusters based on their relative position and velocity and following the agglomerative clustering algorithm. Initially, each tracked point composes its own cluster; then, at each iteration, we merge two clusters with the smallest distance given in the equation:

$$d(c_i, c_j) = \alpha * \Delta V(c_i, c_j) + (1 - \alpha) * \Delta L(c_i, c_j); \qquad (1)$$

where $d(c_i, c_j)$ is the distance measure between two clusters c_i and c_j, $\Delta L(c_i, c_j)$ is the Euclidean distance between the clusters' mean positions, $\Delta V(c_i, c_j)$ is the difference in the clusters' mean velocities, and α is a coefficient giving more importance to one of the characteristics. We set this coefficient to $\alpha = 0.8$ (giving more importance to velocity) by optimizing the proto-objects detection rate (see section 5.2) on a set of objects demonstrations.

We continue to merge GFT points into clusters until a specified threshold on the minimal distance is reached. This threshold is set to 0.0087, also by optimizing the proto-objects detection rate. Each resulting cluster of coherent GFT points is a proto-object and it is the basic element of our following processing. Each detected proto-object is tracked over images considering as tracked from the previous image in the case of tracking more than a half of its GFT points.

Each proto-object can be segmented from the background based on a convex hull of its GFT points. However, this convex hull does not always correspond to the real object boundary. If a convex hull is based on few GFT points, it often cuts the proto-object border or captures the background and surrounding items. In order to improve the proto-object segmentation, the results of tracking performed on RGB images are consolidated with processing of the depth data, and the depth variation in the visual field is used to obtain more precise boundaries.

When processing the depth data, at first, the Median blur filter [29] is applied to smooth depth values and reduce the noise in the data. Then, the Sobel operator [13] based on the first derivative is used to detect horizontal and vertical edges allowing to reveal the depth variation in the visual field. Noisy and non-significant edges are filtered out by thresholding the obtained results, then the dilation and erosion operations [60] are used to close broken contours. The obtained

[2] implemented in the OpenCV library http://opencv.org

continuous contours are transformed into binary masks. An additional interest of this step is its advantage in segmenting several static physical entities localized close to each other; so if a convex hull of GFT points groups together several static entities, the processing of the depth data allows to isolate the corresponding proto-objects inside a single convex hull.

3.2 Entity appearance representation

The appearance of each of the proto-objects regions obtained in the previous section should then be characterized in order to be learned or recognized later on in our system. For this objective, we use complementary low-level visual features that are further organized into hierarchical representations, as shown in Fig. 3. The appearance of a proto-object corresponds to a *view*, i.e., the appearance of an entity observed from one perspective. The view representation is based on the incremental Bag of visual Words (BoW) approach [18] extended by an additional feature layer incorporating local visual geometry. An *entity* then gathers the different appearances of the physical entity in a multi-view model encoded as a set of *views*. Note that a *view* can appear in several *entities* when two different objects share a common appearance from a particular point of view.

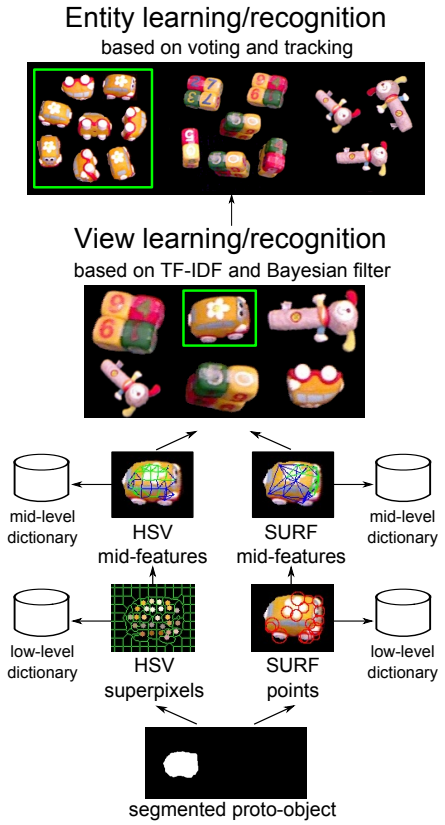


Fig. 3 Construction of an entity representation model.

The robot should be able to deal with various entities, ranging from simple homogeneous objects with few features, to complex textured objects. We choose a combination of complementary visual features that could represent all these objects. As a local descriptor, we use SURF [2] due to its efficient and accurate characterization of local image areas, thus providing a good description of objects with many details. In order to deal with both textured and homogeneous coloured objects, we develop an additional descriptor operating on the level of regularly segmented image regions. The superpixels algorithm [45] is used to segment images into relatively homogeneous regions by grouping similar adjacent pixels. For segmentation, we use the watershed algorithm [4] on the image convolved with Laplacian of Gaussian, initialised with regularly spaced seeds. Each resulting superpixel is characterized by its average color encoded in the HSV space (hue, saturation and value). Note that this segmentation if used to represent a proto-object as a set of colored regions and does not modify the proto-object segmentation obtained in Section 3.1.

The extracted low-level feature descriptors are incrementally quantized into dictionaries of visual words [18]. Starting with a dictionary containing the first feature, each new feature is assigned to its nearest dictionary entry (a visual word) based on the Euclidean distance between their descriptors. If the distance between the current descriptor and each dictionary entry exceeds a threshold, a new visual word is added to the dictionary (see algorithm 1). The quantization procedure provides two dictionaries, one for SURF descriptors and one for superpixel colors. The thresholds for the dictionaries were empirically chosen by optimizing the object recognition rate (see section 5.3) on a small set of representative objects (both textured and textureless).

Data: Feature descriptor
Result: Corresponding visual word
if *Dictionary is empty* **then**
  Add descriptor as the first visual word;
  Return visual word;
else
  dist_min = distance to the nearest visual word;
  if *dist_min* $<$ *threshold* **then**
    return nearest visual word;
  else
    Add descriptor as a new visual word;
    Return new visual word;
  end
end

Algorithm 1: Feature search and dictionary update

The size of the color dictionary remains relatively stable after processing several objects, since colors repeat among different objects quite often. However, the SURF dictionary grows continuously with the number of objects. In order to

avoid the rapid growth of the SURF dictionary, we filtered the SURF features before including them in the dictionary. Only features that are seen over several consecutive frames (we use three consecutive frames) are stored in the dictionary which is used in the following processing as a ground level for view representation.

The low-level features are grouped into more complex mid-level features defined as pairs of low-level features. This feature layer incorporates local visual geometry and allows not only to characterize views by a set of features, i.e. isolated colors or SURF points, but also synthesize information into a more robust description considering relative feature position. For both types of features, each low-level feature is used to construct mid-features with 4 neighboring low-level features[3] that are the closest in terms of the Euclidean distance in the image space. Thus, each mid-feature m_k is a pair of visual words, implicitly encoding the corresponding visual features that have been perceived close in the image space:

$$m_k = (w_a, w_b), \tag{2}$$

where m_k is a mid-feature, w_a and w_b are two visual words corresponding to neighbouring visual features.

Mid-features are incrementally quantized into dictionaries following the same concept used for quantization of low-level features. The dissimilarity measure between two mid-features is estimated as the minimum of pairwise Euclidean distances between their descriptors (eq. 3). The quantization procedure provides dictionaries of SURF-pairs and superpixel-color-pairs.

$$\Delta(m_1, m_2) = min \begin{cases} \Delta F(a_1, a_2) + \Delta F(b_1, b_2), \\ \Delta F(a_1, b_2) + \Delta F(b_1, a_2), \end{cases} \tag{3}$$

where m_1 and m_2 are two compared mid-features, and each mid-feature is a pair of features a and b; ΔF is the dissimilarity between two features (one feature from the first pair and another feature from the second pair).

According to our representation model, all constructed mid-features are used to characterize proto-objects appearances, i.e., views, and each view is encoded by the occurrence frequencies of its mid-features:

$$v_j = \{m_k\}, \tag{4}$$

where v_j is a view and m_k is a mid-feature.

In images captured by a visual sensor, a 3D object is perceived as its 2D projection depending on its position and viewing angle. These projections can differ significantly depending on the object appearance and shape and can also depend on the illumination when reflected light produces shadows and saturations making invisible some parts of the object [22]. In our approach, the overall appearance of each physical entity is characterized by a multi-view representation model (see Fig. 4) that covers possible changes in the appearance of an entity emerging from different viewing angles and varying illumination. Each entity is encoded as a collection of views, where each view characterizes the appearance of one perspective of the entity:

$$E_i = \{v_j\}, \tag{5}$$

where E_i is an entity and v_j is its observed view. Note that one view may be a part of several entities.

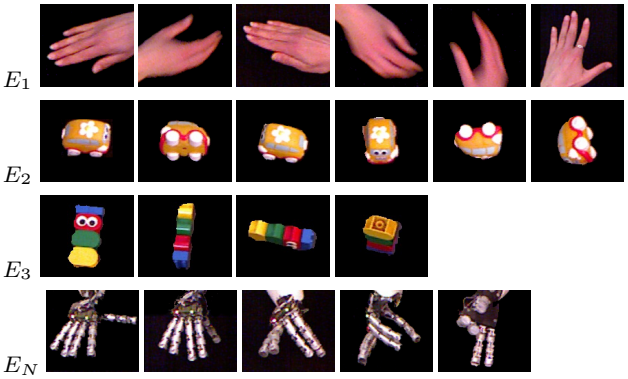


Fig. 4 Examples of representation models of four different entities (each model with its views is shown in one line).

3.3 View learning and recognition

Each proto-object detected in the visual space is either recognized as a known view or learned as a new view. The view recognition procedure consists of a likelihood estimation using a voting method based on TF-IDF (Term-Frequency - Inverse-Document Frequency) [63] approach followed by a Bayesian filter estimating a posteriori probability of being one of the known views.

The voting method (see Fig. 5) is used to estimate the likelihood of a set of mid-features (extracted from the proto-object region) being one of the known views. Each mid-feature quantized into a visual word votes for a view where it has been seen before with its TF-IDF score. The TF-IDF score is aimed to evaluate the importance of visual words with respect to views and give higher weights to distinctive visual words. The voting method is fast, since it uses an inverted index allowing to consider only the views that have at least one common mid-feature with the analyzed proto-object. The advantage of this approach with respect

[3] We tested 2, 3 and 4-connectedness of features and chose 4-connectedness based on our preliminary experiments with a set of 10 objects as a compromise between performance and computational cost in order to be able to perform interactive experiments. We also compared the use of low-level and mid-level features, and got an improvement from 84.33% to 97.83% recognition rate (based on pure labels) when using mid-features. More details can be found in [41], p84.

to supervised algorithms, like Support Vector Machines or boosting, is the ability to learn new views incrementally by updating mid-feature occurrence statistics, without knowing the number of views in advance and without re-processing all the data while adding a new view.

More formally, the likelihood of a mid-feature set $\{m_k\}$ being the view v_j is computed as a sum of products of mid-features frequencies and the inverse view frequency:

$$L(v_j) = \sum_{m_k} tf(m_k) idf(m_k), \tag{6}$$

where $tf(m_k)$ is the occurrence frequency of the mid-feature m_k, and $idf(m_k)$ is the inverse view frequency for the mid-feature m_k.

The occurrence frequency of the mid-feature is computed as:

$$tf(m_k) = \frac{n_{m_k v_j}}{n_{v_j}}, \tag{7}$$

where $n_{m_k v_j}$ is the number of occurrences of the mid-feature m_k in the view v_j, and n_{v_j} is the total number of mid-features in the view v_j.

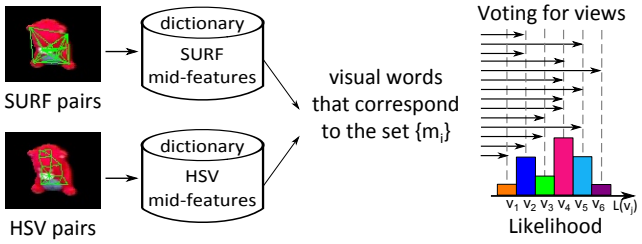


Fig. 5 The voting method: each extracted mid-feature votes for views, where it has been seen before.

The inverse view frequency $idf(m_k)$ is related to the occurrence frequency of a mid-feature among all seen views; it is used to decrease the weight of mid-features, which are often present in different views, and it is computed as:

$$idf(m_k) = log \frac{N_v}{n_{m_k}}, \tag{8}$$

where n_{m_k} is the number of views with the mid-feature m_k, and N_v is the total number of seen views.

The estimated likelihood is used for appearance-based recognition of views. However, views of different objects can be similar, and one object observed from a certain perspective can resemble another object. The recognition becomes even more difficult if an object is occluded that often happens during manipulations. In our approach, the temporal consistency of recognition is improved by applying a Bayesian filter in order to reduce the potential confusion between entities recognized on a short time scale. Based on tracking, we predict the probability of recognizing the view from the a priori probability computed in the previous image and the probability of being tracked from the previous image. The final a posteriori probability of recognizing a view is estimated recursively using its likelihood and its prediction:

$$p_t(v_j) = \eta L(v_j) \sum_l p(v_j|v_l) p_{t-1}(v_l), \tag{9}$$

where $L(v_j)$ is the likelihood of recognizing the view v_j, $p(v_j|v_l)$ is the probability that the current view is v_j if the view v_l was recognized in the previous image (we set $p(v_j|v_l)$ equal to 0.8 if $v_j = v_l$ and $0.2/(N_v - 1)$ otherwise, with the total number of views being N_v), $p_{t-1}(v_l)$ is the a priori probability of the view v_l computed in the previous image, and η is the normalization term.

Depending on the highest a posteriori probability obtained among all known views, the proto-object can be

- stored as a new view with the set of current mid-features, if the highest probability is lower than the threshold $th_{v.n.}$,
- recognized as the view with the highest probability and updated with the current set of mid-features, if the probability is higher than the threshold $th_{v.u.}$,
- recognized as the view with the highest probability but not updated, otherwise.

The thresholds $th_{v.n.}$ and $th_{v.u.}$ allow to perform only stable updates in case of high confidence of recognition and create new views only in case of low probability of recognition, thus allowing to avoid duplicating views in memory. The update of the recognized view consists simply in updating the number of occurrences $n_{m_k v_j}$ and n_{v_j} of each mid-feature in the view and the number of views containing the mid-feature n_{m_k} used for computing the $tf - idf$ score (eq. 7 and 8).

3.4 Entity learning and recognition

The multi-view appearance model of the corresponding entity should finally be updated with the current view. Each identified view is therefore associated with an entity either using tracking, or appearance-based recognition. In the case of successful tracking from the previous image, the current view is simply associated with the entity recognized in the previous image (see Fig. 6). When the entity is not tracked from the previous image, because the entity just appeared or because of tracking failure due to motion blur for example, the entity is recognized using a maximum likelihood approach based on a voting method similar to the one used for recognizing views.

The likelihood of the view v_j being a part of one of already known entities is computed as:

$$L(E_i) = tf(v_j) idf(v_j), \tag{10}$$

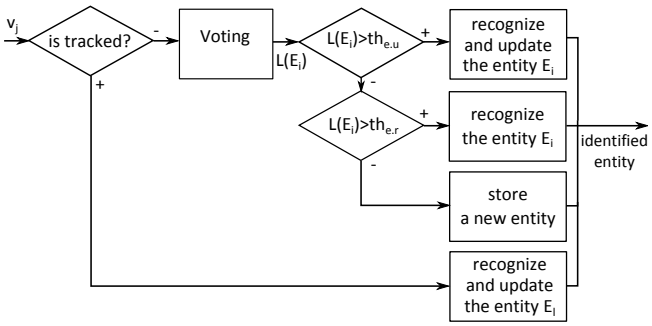


Fig. 6 The main steps of the entity learning/recognition; where v_j is the current view, E_i is the entity corresponding to v_j with the maximal likelihood $L(E_i)$, E_l is the entity tracked from the previous image.

where $tf(v_j)$ is the occurrence frequency of the view v_j, and $idf(v_j)$ is the inverse entity frequency for the view v_j.

The occurrence frequency of the view is computed as $tf(v_j) = \frac{n_{v_j E_i}}{n_{E_i}}$, where $n_{v_j E_i}$ is the number of occurrences of the view v_j in the entity model E_i, and n_{E_i} is the number of views in the entity model E_i.

The inverse entity frequency is related to the view occurrence among all entities; it is used to decrease the weight of views, which are often present in models of different entities: $idf(v_j) = log\frac{N_E}{n_{v_j}}$, where n_{v_j} is the number of entities with the view v_j, and N_E is the total number of seen entities.

The entity recognition decision is based on several thresholds (similar to the recognition of views). The entity can be

- stored as a new entity with the current view, if the maximal likelihood is lower than the threshold $th_{e.n.}$,
- recognized as the entity with the maximal likelihood and updated with the current view, if the likelihood is higher than the threshold $th_{e.u.}$;
- recognized as the entity with the maximal likelihood but not updated, otherwise.

By identifying physical entities and tracking them over time, their multi-view representation models (see Fig. 4) are constructed and updated with the observed views.

3.5 Connected entities recognition

In our scenario, objects are explored through manipulation. As we have observed during our experiments, object manipulation introduces additional difficulties in processing of the visual data: both the hand and the grasped object are detected inside a single proto-object and moreover, a hand holding the object produces multiple occlusions and sometimes divides the grasped object into parts. Therefore, we process each proto-object in a way allowing to recognize it as several connected entities. This problem requires object segregation, as it is called in psychology. The object segregation capability is an important aspect of our approach which is capable of segmenting connected entities based on already acquired knowledge about entities seen alone.

In our approach, each proto-object is recognized either as a single entity or two connected entities based on the following double-check procedure (see Fig. 7):

1. all mid-features of the proto-object are used for recognition of the most probable view among all known views, as described in Section 3.3,
2. the mid-features that do not appear in the most probable view are used for recognition of a possible connected view using the same procedure. The connected view is recognized if its probability is higher than $th_{v.c.}$. If this recognition probability is low and more than 20% of mid-features do not correspond to the first recognized view, then a new view is stored with these mid-features.

Fig. 7 Connected entities recognition: a)all extracted mid-features (HSV pairs); b)the mid-features of the first recognized view, c)the mid-features of the first recognized view (shown by pink color) and the mid-features of the connected view (shown by blue color).

Further, each identified view is associated with one of physical entities as described earlier. If both the object and the hand have been already seen separately, the corresponding entities exist in the visual memory, and they can be recognized as connected entities.

The ability to recognize connected entities is really important in scenarios with object manipulation. It helps preventing erroneous updates of views and entities models when the object is grasped. If both the object and the hand are identified as connected entities, then the view of the object will not be updated with the mid-features of the hand. Furthermore, the information about connected entities is also used during the entity categorization and interactive object learning presented in the following section.

4 Interactive learning

In this section, we describe our second developmental stage where the robot manipulates objects to improve their models. As a pre-requisite, our approach first categorize all entities into parts of the robot, parts of a human partner, and manipulable objects. This categorization process makes our approach robust to changes in the robots effector appearance and allows to update objects models efficiently without adding parts of the robot or human hands to the objects

models. Note that all the processes described in the previous sections are still active, thus making it possible to interleave the two developmental stages by introducing new entities at any time.

4.1 Entity categorization

The categorization procedure is aimed at identifying the nature of physical entities detected in the visual field in the interactive scenario when the robot and a human partner manipulate objects. Each physical entity is classified into one of the following categories: a part of the robot c_r, a part of a human partner c_h, an object c_o, an object grasped by the robot c_{o+r}, or an object grasped by a human partner c_{o+h}. Before identification of the body of the robot, which is a requirement for the identification of other categories, all entities are temporally associated to the unknown category c_u, and their correct categories are identified later. Therefore, within the categorization procedure, at first, the parts of the body of the robot are discriminated among all physical entities, and then, the rest of single entities are distinguished either as a human part, or a manipulable object category, as shown in Fig. 8.

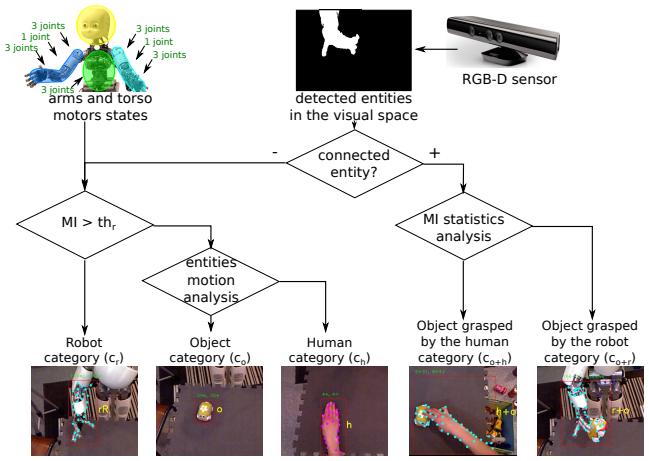


Fig. 8 The categorization flowchart: parts of the robot c_r are discriminated based on mutual information (MI) between the visual and proprioceptive data; parts of a human partner c_h and objects c_o are distinguished based on both MI and statistics on entities motion; connected entities are categorized either as an object grasped by the robot c_{o+r} or an object grasped by a human partner c_{o+h}.

4.1.1 Robot self-identification

Our goal is to implement a strategy that requires minimum prior knowledge and avoids the need for a predefined appearance of the robot, a joint-link structure, or a predefined behaviour. The independence on the appearance should allow a robust recognition of the hands of the robot in the case of changing their appearance, in the case of occlusion, and in the case of extension of the hands by grasped tools. The independence on the behaviour enables to perform recognition at any time, during a variety of interactive actions without requiring a specific identification phase.

Therefore, during the motor activity of the robot (the actions performed by the robot will be described in Section 5.1.3), the visual information is gathered together with the proprioceptive data, and based on mutual information (MI) between these senses, the system identifies the parts of the body of the robot among detected physical entities. As the input data, we acquire and process:

- visual information: the position of detected entities in the visual field,
- proprioceptive information: joints values of the robot's motors accessed through YARP ports[4]:
  - arms joints: shoulder (pitch, roll, and yaw), elbow, and wrist joints (pronosupination, pitch, and yaw),
  - torso joints: pitch, roll, and yaw.

The data acquisition is driven by the visual perception, and the states of the motors are acquired after receiving a new image from the visual sensor. The motors states are acquired as a set of arm-torso joints values without considering the functionality of each joint, nor the character of its impact on the displacement of the hands of the robot. We acquire one set of joints values per arm with the torso joints in each set. The head motors however are not analysed, since they do not effect on the position of the hand observed from our external visual sensor.

Both visual and proprioceptive data need to be quantized in order to compute mutual information. For the visual space being only of dimension 2, a simple regular grid is used: for each detected entity, its position in image space is quantized into one of the visual clusters obtained by dividing the image space with a regular grid of 12 columns and 10 rows, producing 120 rectangular visual clusters. The joint space however has a higher dimensionality, and it is not possible to use a regular discretization along all the dimensions. The joints values are therefore quantized into a dictionary of arm-torso configurations with each entry encoded as a vector of joints values. The quantization is incremental, i.e., it adds new clusters as required by the data, and we use the same algorithm as for visual dictionary creation (algorithm 1). This leads to a sparse representation of the joint space that will adapt to any new joint configuration experienced by the robot. In our experiments, the mean number of arm-torso configurations generated by this procedure was about 40.

Mutual information is used to evaluate the dependencies between the arm-torso configurations A_k (either left or right arm) and the localization of each physical entity E_i in the

[4] http://eris.liralab.it/yarp

visual cluster L_{E_i}:

$$MI(L_{E_i}; A_k) = H(L_{E_i}) - Hc(L_{E_i}|A_k), \quad (11)$$

where L_{E_i} is the position of the entity quantized into the visual cluster, A_k is the state of the arm k of the robot quantized into the arm-torso configuration, $H(L_{E_i})$ is the marginal entropy, and $Hc(L_{E_i}|A_k)$ is the conditional entropy computed in the following way:

$$H(L_{E_i}) = -\sum_l p(L_{E_i} = l) log(p(L_{E_i} = l)), \quad (12)$$

$$Hc(L_{E_i}|A_k) = -\sum_a p(A_k = a) \times \sum_l p(L_{E_i} = l|A_k = a) log(p(L_{E_i} = l|A_k = a)) \quad (13)$$

where $p(L_{E_i} = l)$ is the probability that the localization of entity E_i is the visual cluster l; $p(A_k = a)$ is the probability that the arm-torso configuration A_k is the configuration a, and $p(L_{E_i} = l|A_k = a)$ is the probability of the entity being in the cluster l when the arm k is in the cluster a.

While the robot moves its hands in the visual field, gathering statistics about the localization of entities and the occurrences of arm-torso configurations, MI grows for the physical entities that correspond to the hands of the robot. When the MI value reaches a specified threshold th_r, the entity is identified as the robot category c_r. On the contrary, the human and the object categories should have smaller MI due to their independence from the motors of the robot. The threshold for identifying the robot category was empirically chosen based on MI distribution obtained on a small labelled set of robot and non-robot entities. Thereby, the physical entity is identified as the robot category c_r, if its MI is higher than th_r, and otherwise, it is considered as one of the non-robot categories that will be identified in the following processing as described in the next subsection.

In case of changing the appearance of the hand of the robot (like wearing gloves, that we do during our experiments presented in Section 5.4), the robot category c_r can be associated with several entities with each entity characterizing a different appearance of the hand (see Fig. 9).

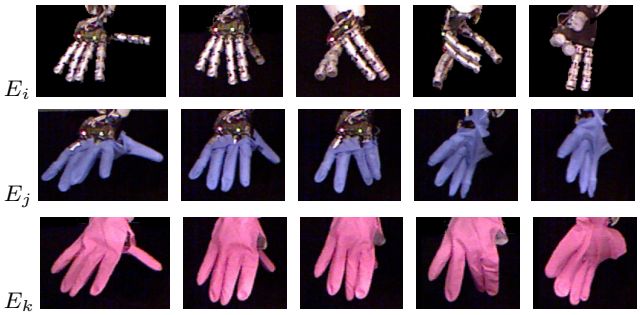


Fig. 9 Examples of multi-view models of three entities characterizing different appearances of the hand of the robot (each model with its views is shown in one line).

4.1.2 Discrimination of manipulable objects and human parts

Among the non-robot physical entities, human parts and objects are discriminated based on their motion behaviour. Most objects, like the one used in our experiments, are static most of time, and they are displaced by the robot or its human partner. Among categories analyzed in this work, only the robot and the human categories can move alone (while not connected to other entities). Thus, we accumulate the statistics on entities motion and use it to distinguish the object category as a mostly static entity that moves only when connected to other entities (see algorithm 2). Note that this definition is linked to our scenario and is not universal: we would recognize objects moving autonomously or animals as human parts, while a human moving his left hand only when it is touched by his right hand would see the right hand categorized as a human part and the left hand as an object.

While detecting physical entities, we accumulate the statistics on their motion over time. Based on these statistics and the output from the self-identification algorithm, the following probabilities are estimated:

- $p_s = p(E_i|c_{E_i} \neq c_r)$ the probability of seeing the entity E_i moving as a single entity while being identified as a non-robot category,
- $p_c = p(E_i|c_{E_i} \neq c_r, c_{E_{i2}} = c_r)$ the probability of seeing the entity E_i identified as a non-robot category and moving together with the connected entity E_{i2} identified as a robot category.

Analysing the motion statistics of single entities, the probability p_s should be lower for the object category, since object entities usually do not move alone, as discussed earlier. Analysing the motion statistics of connected entities, the probability p_c should be higher for the object category, since object entities often move together with other entities for example, when objects are manipulated. Thereby, each non-robot entity is categorized as:

- the object category c_o, if the probability $p_c > th_o.c.$ and $p_s < th_o.s.$;
- the human category c_h, otherwise.

Following our approach, the parts of the body of the robot are identified first, so that before the robot starts interaction with objects it has already accumulated some statistics on entities motion. Once the robot starts interaction with objects, it accumulates statistics on motion of entities together with its hands. While applying our categorization algorithm to each detected entity, we identify each single entity as one of the following categories: c_o, c_h, or c_r (see

Data: Non-robot entity E_i
and optionally its connected entity E_{i2}
Result: Category assigned to the entity E_i
if $E_{i2} = \emptyset$ **then**
  update the probability of moving alone p_s;
else
  if *category of* $E_{i2} = c_r$ **then**
    update the probability of moving when connected p_c;
  end
end
if $(p_c > th_{o.c.})$ *and* $(p_s < th_{o.s.})$ **then**
  assign the object category, c_o, to E_i;
else
  assign the human category, c_h, to E_i;
end

Algorithm 2: Discrimination of manipulable objects and human parts

Fig. 10). Connected entities are identified either as an object grasped by a robot category c_{o+r} or an object grasped by a human category c_{o+h} based on the categorization statistics gathered when the corresponding entities have been seen alone.

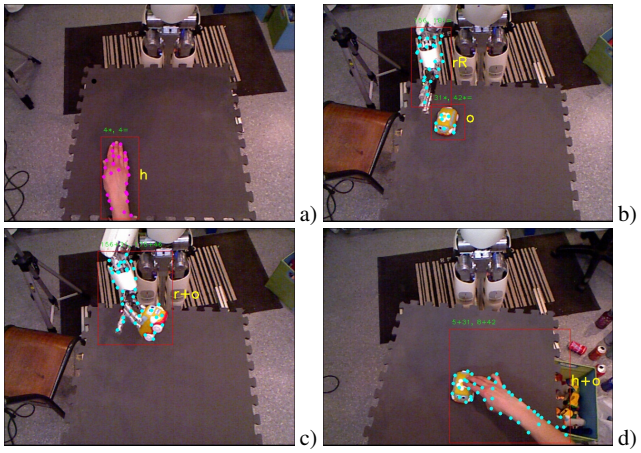


Fig. 10 Entity categorization examples: a)the human hand identified as c_h; b) the hand of the robot identified as c_r and the object identified as c_o; c) the object grasped by the robot identified as c_{o+r}; d) the object grasped by the human identified as c_{o+h}.

4.2 Interactive object learning

Once the robot is able to detect and categorize physical entities in the visual space, it starts to interact with object entities (see Fig. 15 and Fig. 16). The actions executed on the robot are described in Section 5.1.3. While interacting with an entity, the system each time remembers the grasped entity as E_g and the model of this entity is updated during the action of the robot. This is a kind of self-supervision, where the object is supposed to remain the same during its manipulation.

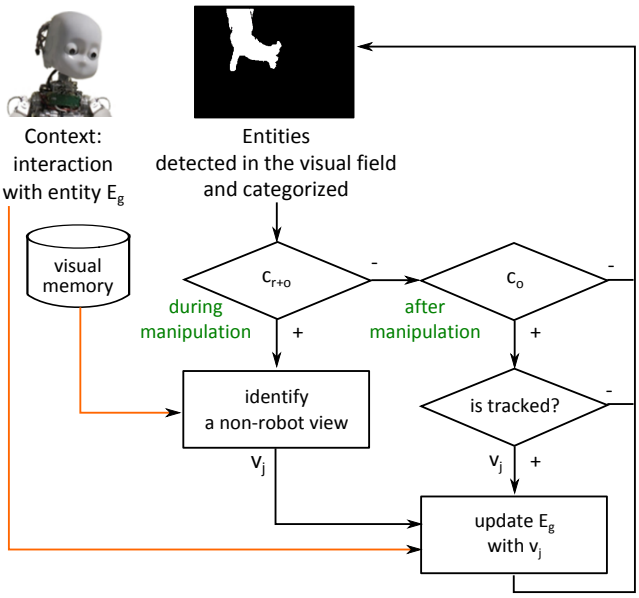


Fig. 11 Improving the object representation model during interaction. During the action of the robot, the manipulated entity E_g can be detected either as an entity connected to the hand of the robot and identified as the *object+robot* category c_{o+r}, or as a single entity identified as the *object* category c_o. In both cases, the manipulated entity E_g can be updated with the non-robot view v_j recognized in the current image (see text for details).

According to our algorithm, the system continuously detects entities in the visual space and categorizes them. While the robot interacts with an object, we are able to discriminate between the object entity and the robot entity, when they move separately or together (e.g. when the object is grasped). The information about identified categories of entities is used by our interactive learning algorithm summarized in Fig. 11. If during interaction with an object, the system detects connected entities categorized as *the object grasped by the robot*, we verify the categories of connected views. For this purpose, we retrieve the set of entities $\{E_i\}$ with the current view in their models. For each entity, we retrieve its category $\{c_{E_i}\}$ from the statistics stored in the memory, and based on these categories the view is identified as:

- a robot view, if at least one corresponding entity is identified as the robot category ($\exists i, c_{E_i} = c_r$);
- a non-robot view, if none of corresponding entities is identified as the robot category ($\forall i, c_{E_i} \neq c_r$).

If connected views are identified as a robot view and a non-robot view (see Fig. 12), the model of the grasped entity E_g is updated with the non-robot view.

While finishing the object manipulation process, the robot releases its hand, and the grasped object falls down. In this case, if the object is detected as a single entity with a unknown view (corresponding to a perspective that was not yet observed), a new view will be stored in the memory.

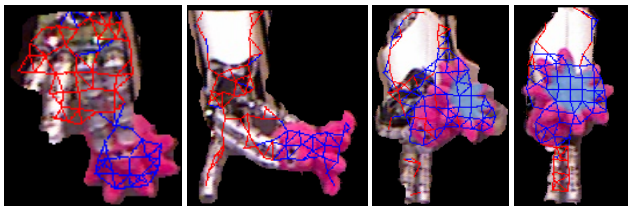


Fig. 12 Examples of connected views and their mid-features (HSV pairs) during interactive object learning: the red mid-features correspond to one of connected views (in this case, the hand of the robot), and the blue mid-features correspond to another connected view (in this case, the object).

Fig. 13 The 20 objects used in our experiments. The objects are numbered from 1 to 20, from top left to bottom right, and this order is preserved in the reported experiments. These images are the real images acquired by the Kinect sensor and used by our system.

The model of this entity could be updated with this new view based on tracking in the following images. Thereby, the robot can explore an object appearance by grasping and throwing it, while updating the model of the manipulated entity with the observed views.

After manipulations with objects, the system performs a check of the visual memory and cleans the dictionaries of entities and views. The entity dictionary is cleaned by suppressing the noisy entities that have no proper views (these entities have only views common with other entities). The view dictionary is cleaned by suppressing the views that have no associated entities; such views could be created during interaction with an entity but never added to its model. Finally, the cleaning of dictionary makes the knowledge about physical entities more coherent and improves object recognition as shown in the next section.

5 EXPERIMENTAL EVALUATION

The proposed perceptual system is evaluated on the iCub[5] (see Fig. 14b) and the Meka[6] (see Fig.14a) humanoid robots exploring their environment in interactive scenarios. Precisely, all quantitative data reported in this paper were acquired on the iCub robot, in its first version [48], with a mean frame rate of 10Hz. In our experiments, at first, the robot learns about its close environment through observation, while a human partner demonstrates objects to the robot, and then, the robot explores its close environment and surrounding objects through interaction. First actions of the robot are aimed at identifying the parts of its own body, then it discriminates manipulable objects and parts of human partners. Once the robot is able to categorize the entities in its visual field, it starts learning objects appearances through manipulation.

The whole set of objects used in our experiments is shown in Fig. 13. We choose both simple homogeneous objects (like toys) and also more complex textured objects (like everyday products including bottles and boxes).

[5] http //www.icub.org

[6] http //en.wikipedia.org/wiki/Meka_Robotics

5.1 Experimental setup

In our setup, the robot is placed in front of a table, and the visual input is taken from the external Kinect sensor mounted above the head of the robot, as shown in Fig. 14. In case of using an external visual sensor, interaction with entities requires their localization not only in the image space but also with respect to the robot. Therefore, at the beginning of our experiments, the visual sensor is calibrated with respect to the robot. During experiments, each detected entity is localized in the operational space of the robot and characterized by its orientation and size.

5.1.1 Visual sensor calibration

The calibration of the visual sensor relative to the base of the robot is performed with a calibration pattern, a chessboard, and the OpenCV library is used to compute the position of the chessboard relative to the sensor. The computation of the transformation matrix requires both the position and orientation of the chessboard in the operational space of the robot. The orientation of the chessboard is known, since it is placed horizontally in front of the robot. In order to obtain the position of the chessboard, we place the hand of the robot above the origin of the chessboard (see Fig. 14) and acquire the position of the hand. Then, the transformation matrix is computed in the following way:

$$T_{sensor \rightarrow robot} = T_{sensor \rightarrow chessbrd} \times T_{chessbrd \rightarrow robot}. \quad (14)$$

5.1.2 Entity localization

For each detected entity, its 3D position in the visual space is estimated with respect to the sensor by processing the RGB-D data as a point cloud and computing the average position of its 3D points. The orientation of the entity is estimated based on eigenvectors and eigenvalues of the covariance matrix of the points. The eigenvectors correspond to three orthogonal vectors oriented in the direction maximizing the variance of the points of the entity along its axis. The eigenvectors are used as the reference frame of the entity. A quaternion is chosen to represent the orientation of

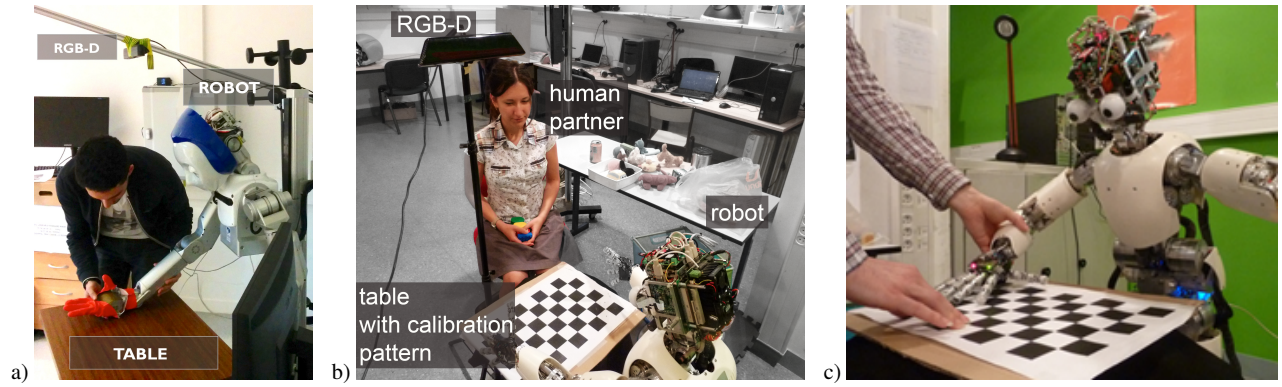


Fig. 14 a) The experimental setup for the Meka robot with the relative position of the sensor, the robot, and the table. b) The experimental setup for the iCub robot. c) The acquisition of the position of the pattern in the operational space of the robot, shown for the iCub robot.

the entity, since this representation is compact, fast, and stable [19]. The position and orientation of the entity is then estimated in the reference frame of the robot using the transformation obtained through the calibration and the *Eigen*3 library[7].

5.1.3 Actions

The interactive actions of the robot are aimed at achieving two main goals: categorization of entities (including self-identification and discrimination of manipulable objects) and learning objects appearances. Both simple action primitives and more complex manipulations have been implemented and used in [31,33]. In this paper, we use two complex manipulations aimed at observing an object from different viewing angles and at different scales:

- *TakeLiftFall* manipulation (see Fig. 15) consists of reaching an object from above, taking it with a three finger pinch grasp, lifting, and releasing. This action generates a random view of the object, when the object falls on the table,
- *TakeObserve* manipulation (see Fig. 16) consists of reaching an object from above, taking it with a three finger pinch grasp, turning the object and approaching towards the camera, and returning it back to the table. This action allows to observe several object perspectives from different viewing angles and also at a closer scale.

These "complex" manipulations are encoded as sequences of simple "atomic" action primitives, such as *reach* or *grasp*. Based on the current state of an object (i.e., its position on the table) and the robot (i.e., the position of its hands and its joints values), actions could have different durations and the speed of fingers movements. In order to grasp an object, the robot approaches its hand towards the top of the object, estimated by the visual system, and executes a three-finger pinch grasp from top. The grasp is pre-encoded and it is designed to be robust for different kinds of objects. As the fingers are tendon-driven, the grasp is naturally compliant, adapting to the shape of the object. Once the object is grasped, the robot continues the sequence of actions to execute the required manipulation. [8]

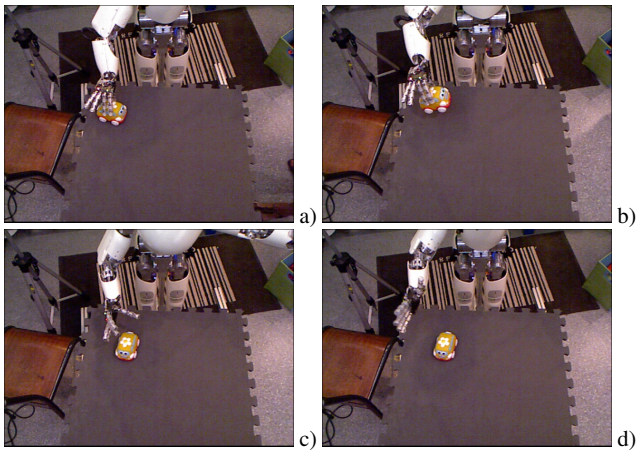


Fig. 15 *TakeLiftFall* manipulation: the object is a)grasped, b)lifted, and c)released; d)when the object falls on the table, it makes it turning into a random perspective.

5.1.4 Evaluation methodology

Since our work is aimed at interactive learning about the close environment of the robot, it makes difficult to evaluate the learning performance using existing image databases.

[7] http://eigen.tuxfamily.org

[8] The code used in these experiments is open-source. Details for installing the code are available at `http://eris.liralab.it/wiki/UPMC_iCub_project/MACSi_Software` while the documentation for running the experiments is at `http://chronos.isir.upmc.fr/~ivaldi/macsi/doc/`. The experiments can be directly reproduced with an iCub robot, whereas in case of other robots with a different middleware (i.e., not based on YARP), the module encoding the action primitives has to be adapted.

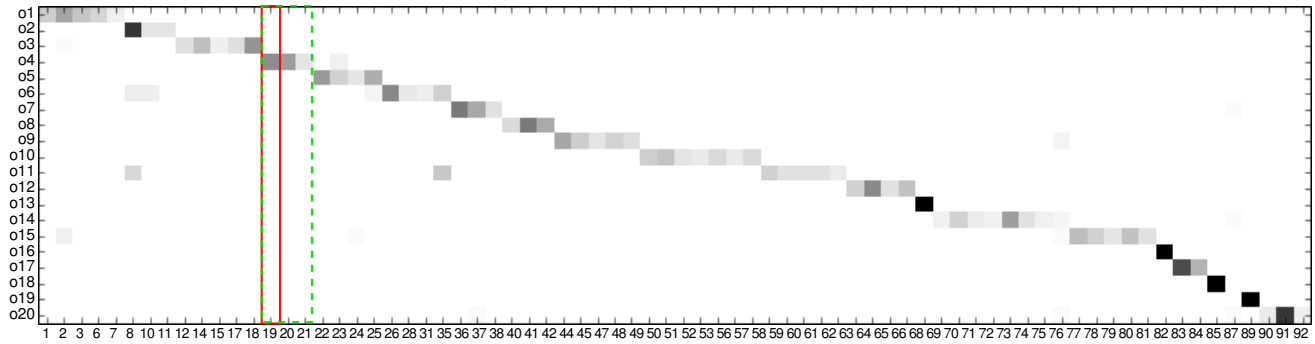


Fig. 17 The association matrix obtained for the 20 objects (shown in rows) and the corresponding physical entities (shown in columns); the color range (from while 0% to black 100%) represents the percentage of object instances associated with each entity; the columns are sorted by the order of created entities that nearly follows the order of learned objects. Among entities associated with each object, we distinguish one major entity that was the most frequently associated (for example, the entity 19 for the object o_4, shown in red solid line) and pure entities that were associated with one object, but never with other objects (for example the entities 19, 20, and 21 for the object o_4, shown in green dashed line).

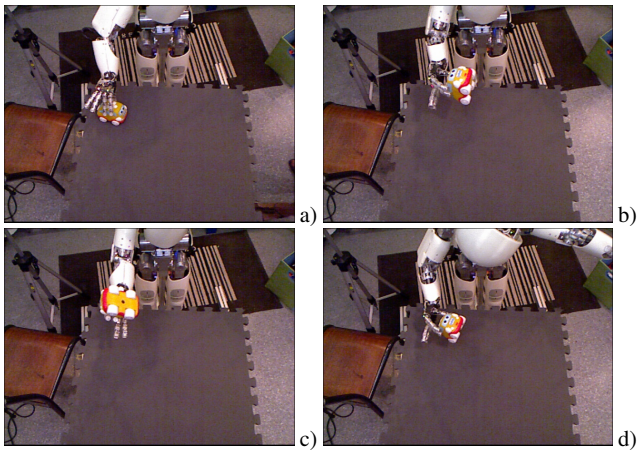


Fig. 16 *TakeObserve* manipulation: the object is a)grasped, b)lifted and approached to the camera, c)turned around, and d)returned back to the table.

Moreover, as learning is incremental and iterative, it is difficult to have a precise evaluation of the performance at a given time during real-time operation. Thus, the performance is evaluated on several stages of developmental learning, and the evaluation is based on pre-recorded sequences of images labelled with a reference ground truth. The evaluation procedure includes estimation of the following characteristics:

- detection rate is obtained based on manually labelled images,
- categorization rate: self-identification is evaluated using forward kinematics model as a reference, while discrimination of objects and human parts is evaluated based on manually labelled images with the correct categories,
- recognition rate is obtained using a separate evaluation image database.

In order to evaluate the object recognition, we make a database with 50 images for each object used in the experiments, and each object is shown from different perspectives. During evaluation, the perceptual system assigns the images of objects from the database to physical entities, and then, we compute the number of entities and views assigned to each real object. The object recognition rate is estimated based on the following entities chosen for each object:

- a major entity as the most frequently associated entity among all entities associated with this particular object,
- pure entities as the entities associated with this particular object, but never with other objects.

Examples of major and pure entities are illustrated in the association matrix in Fig. 17. The object recognition rate is computed as a percentage of the object instances associated with its major/pure entities, with respect to the total number of images with the object.

For all the thresholds used in our algorithms (sections 3.4, 4.1, and 4.2), we ran a first experiment with 10 objects and the initial appearance of the robot and experimentally varied the thresholds to optimize the recognition rates and the categorization performance. We then kept these thresholds for all reported experiments.

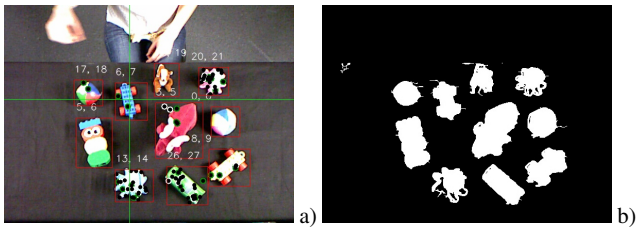
5.2 Evaluation of detection and tracking

In this experiment, the robot learns about its close environment through observation, while a human partner demonstrates the 20 objects (see Fig. 13) one by one. Each object is manipulated for about one minute (that corresponds to about 600 images) allowing to observe different perspectives of the object. In total, the experiment lasts about 20 minutes and contains about 12000 images.

The object detection rate is estimated as a percentage of images with segmented objects, with respect to the total number of images with the object. On average, our system shows an object detection rate of 98% in case of segmenting entities based on depth-contours. We have also compared the detection rate with and without using the depth data. Using motion only (without using the depth data) we obtained a

detection rate of 86%, showing that our system could also work using embedded cameras with a loss of performance.

The tracking rate is estimated as the percentage of tracked instances of the object with respect to the total occurrence of the object in consecutive images. On average, our system shows a tracking rate of 77% that does not depend on the use of the depth information. Note that tracking failures mainly happen with few objects (O_1, O_3, O_{10}, and O_{15}) that have only few features.



a) b)

Fig. 18 Simultaneous processing of several objects: a)10 objects detected and recognized in the visual space of the robot, b)the resulted segmentation of the objects.

5.3 Evaluation of learning through observation

Using the same experiment as in the previous section (i.e., after observing each object for 1 minute), the object recognition rates computed on the separated evaluation database are reported in Table 1. The average recognition rate based on pure entities (i.e., the set of all entities associated only with one object)is 85,7%. The average recognition rate based on major entities (i.e., the pure entity the most frequently associated with the object) is 55,8%. The obtained recognition rates differ between objects. Intuitively, objects with different appearances have been recognized better than objects which are similar to each other. From the association matrix (see Fig. 17), the maximal confusion has occurred between the objects O_{11} and O_6, which have similar colors and similar lego-parts. However, the two identical objects O_1 and O_3 which differ only by color, have been distinguished rather well.

The objects of our dataset that show lower tracking rates (O_1, O_3, O_{10}, and O_{15}) also show smaller recognition rates based on major entities (see Table 1, column 2) comparing to other objects. This is caused by the fact that a tracking failure often leads to the creation of a new entity and prevents to associate several views to a single entity.

From Table 1 and Fig. 17, most objects have been associated with several entities, with an average of 4.1 entities per object. This is a common limitation of unsupervised learning approaches, where the robot decides itself if it observes a new object or a known object. We will see that interactive learning makes it possible to reduce this segmentation of objects into several entities.

We also evaluate our system for simultaneous processing of multiple objects in a single image. The system has been tested with up to 10 objects demonstrated at the same time (see Fig. 18), and all objects have been detected and recognized.

During our experiments on object learning, the average processing time was 0.13s for images with one object. The time required to process one object varies significantly between objects and it depends on their complexity and the number of extracted features. Among all processing stages, the highest computation cost belongs to recognition and learning of views, and in particularly to searching features in dictionaries. Moreover, this cost increases with the dictionaries growth which was observed to be approximately linear in our experiments. Other processing stages (object detection, segmentation, feature extraction, tracking, and categorization) take all together about 0.06s per image, and this processing cost stays relatively stable over time.

5.4 Evaluation of entity categorization

The categorization performance is evaluated in the interactive scenario where both the robot and the human partner perform actions aimed at exploration of the objects close to the robot.

5.4.1 Evaluation of self-identification

In this experiment, the iCub robot performs free hand motion and interactive actions described in Section 5.1.3, while the human partner also moves its hands in the visual space. In total, the experiment lasts about 12 minutes and contains 7200 images. The identification of the body parts of the robot was evaluated using forward kinematics model as a reference. Our approach was evaluated with the robot normal hand appearance and also while changing its appearance by wearing coloured gloves (see Fig. 9). The categorization procedure was able to identify the hand appearances after a duration varying between 5 and 12 seconds of their motion in the visual field ($c_u - c_r$ in Fig. 19), corresponding to the processing of between 50 and 120 images (see Fig. 19). These variations depend on the particular motions performed by the robot: motions of the hand across the whole visual field are more informative than motions that produce little visible variations and therefore lead to a faster increase of mutual information and a faster hand identification. Once the hand of the robot was first identified, the system has shown an average self-recognition rate of 98.2% for the initial appearance of the hand. The self-recognition rate for the other appearance was 98.1% for the blue glove and 98.0% for the pink glove. Similar results confirming the indepen-

dence of our approach on the hand appearance were obtained with the Meka robot wearing coloured gloves.

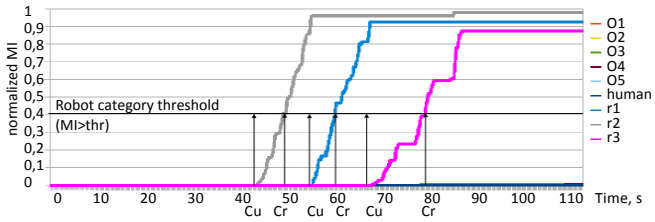


Fig. 19 Categorization of entities performed while both the human partner and the robot (with three different appearances of the hand) perform free hand motion and the human partner also interacts with first five objects: the graph shows the normalized MI value for each entity; each entity appears in the timeline as an unknown category c_u, and once it is categorized, its category is marked in the timeline (in this case, the category c_r). The curves corresponding to the five objects do not appear in the graph as their probability remains close to 0 and they are hidden by the curve corresponding to the human.

5.4.2 Evaluation of categorization of objects and human parts

Once the robot identifies its hands among the physical entities detected in the visual field, it continues interactive exploration of other entities. While both the robot and its human partner perform interactive actions with the objects, the perceptual system continuously analyses the entities behaviour and categorizes them. In total, this experiment lasts about 60 minutes and contains about 36000 images, where the human manipulates each of 20 objects (in total, about 20 minutes), and the robot manipulates each of 20 objects (in total, about 40 minutes). The ability to discriminate the objects and human parts is evaluated a posteriori based on images labelled with the correct entities categories. During the experiment, each object has been successfully identified in the object category within 5-10 seconds of motion during interaction (corresponding to 50-100 images), leading to a total correct categorisation rate of 84%. Human parts have been categorized correctly in 89% of all images. Fig. 20 shows the evolution of the probability of each non-robot entity being an object. It also shows the probability of being a human, given that the two probabilities sum to 1.

5.5 Evaluation of interactive object learning

Once the robot is able to categorize physical entities detected in the visual field, its focuses on interactive object exploration. The robot manipulates each object following the *TakeLiftFall* or *TakeObserve* schemes, described in 5.1.3. Each manipulation lasts about one and a half minute (corresponding to about 900 images). In total, the experiment lasts

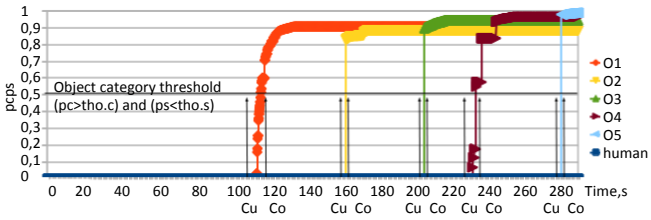


Fig. 20 Categorization of entities performed while the robot interacts with the first five objects: the graph shows the probability of being in the object category based on p_c and p_s for each entity. Each entity appears in the timeline as an unknown category c_u, and once categorized as an object is marked c_o. The entities with the probability below the threshold fall in the human category.

about 30 minutes for each type of manipulation and contains about 18000 images. The performance of interactive learning is evaluated using the database described in Section 5.1.4. The evaluation results are reported in Table 1, where each value is presented in a pair with the corresponding result obtained during learning through observation presented in Section 5.3.

Table 1 Performances of object learning: each value is presented in a pair comparing the results of learning through interaction (2nd stage) / with respect to learning through observation (1st stage)

Object	Recognition rate based on pure entities,%	Recognition rate based on a major entity,%	Number of pure entities	Number of pure views
O_1	96/96	45/33	4/6	9/9
O_2	100/90	92/78	3/3	8/6
O_3	98/96	82/40	3/6	5/6
O_4	58/60	44/44	1/3	2/4
O_5	91/41	52/41	3/1	3/2
O_6	63/63	40/40	4/7	4/7
O_7	60/60	60/52	1/2	1/2
O_8	100/100	86/50	3/4	4/4
O_9	89/96	33/32	4/8	5/9
O_{10}	80/80	23/22	5/8	5/8
O_{11}	84/84	35/23	5/6	6/6
O_{12}	87/87	63/47	2/4	2/4
O_{13}	100/100	100/97	1/2	2/2
O_{14}	87/87	51/38	4/7	4/7
O_{15}	94/90	41/25	3/5	3/5
O_{16}	100/100	100/100	1/1	1/1
O_{17}	100/100	100/80	1/2	2/2
O_{18}	100/100	100/99	1/2	1/2
O_{19}	100/100	100/99	1/1	2/2
O_{20}	83/83	76/76	2/4	2/4
Mean	88.5/85.7	66.2/55.8	2.6/4.1	3.6/4.6

For most of objects, the interactive learning shows an improvement of the recognition rate based on a major entity with respect to the results of learning through observation (see Fig. 21). The recognition rate based on pure entities remains nearly stable in comparison to learning through ob-

servation. These results can be explained by the concept of the learning algorithm aimed at updating the best model of a grasped entity during its manipulation. Thus, interactive learning procedure improves mostly the major entity, while leaving other pure entities without significant changes.

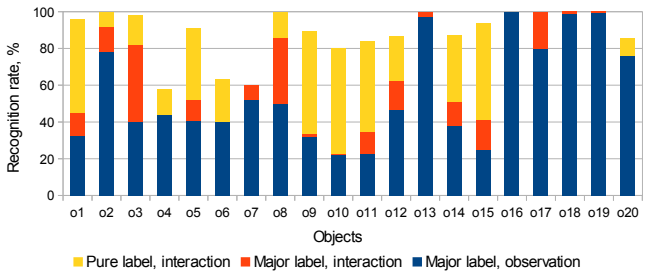


Fig. 21 Improvement of the object recognition rate: the recognition rate (based on major entities) obtained during learning through observation is shown in blue, its improvement during interactive learning is shown in orange, and the final recognition rate (based on pure entities) is shown in yellow.

Interactive learning allows to obtain enhanced objects models with an increased number of views. For objects whose appearances significantly vary between perspectives, interactive learning is especially useful. While manipulating an object, the perceptual system integrates the recognized views into the representation model of the entity thus enhancing the model and making it more complete. Moreover, the system creates new views when it observes previously unknown perspectives of the object. From our experiments, interactive learning results in enhancement of the entities models of the objects O_1, O_2, O_3, O_8, O_9, and O_{11}. The examples of improvements of some models (in particularly, the views added to these models) are shown in Fig. 22.

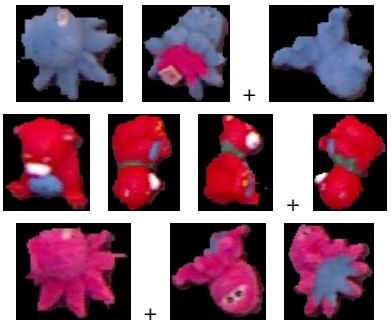


Fig. 22 The representation models of the major entities of the objects O_1, O_2, and O_3 (each model with its views is shown in one line), where the views added during interactive learning are shown after the $+$ sign.

As discussed in Section 5.3, learning through observation results in association of some objects with several physical entities. However, interactive learning allows to consolidate the knowledge about an object within its major entity and decrease the number of entities associated with the object. The total number of entities and views decreases mostly due to cleaning dictionaries performed after manipulation and described in Section 4.2. Cleaning dictionaries makes the knowledge more coherent by removing noisy entities and thus leading to the improvement of the object recognition rate based on major entities as less views are associated to the noisy entities.

6 Discussion

We have evaluated our system with a set of objects varying in color and texture, showing its ability to integrate both information for recognition, and its capability to recognize and learn object even when manipulated. However, the choice of the bag of word approach for object representation and hand-crafted features could probably be improved, for example using even more geometric information than we have used in feature pairs. Another interesting approach would be to learn the visual features themselves, which proved to be efficient in a number of applications [39]. Regarding the kind of objects our system can learn, our multi-view model should be well adapted to objects changing shapes, such as articulated objects. The different appearances corresponding to the change in the articulated objects would be integrated as other views, as long as object tracking is possible during object modification.

From a computational point of view, scaling our approach to a larger set of objects will face the issue of feature dictionary growth (Section 5.3) that increases the view learning and recognition time. In our system, the mean computation time for 20 objects is 0.07s for view learning and recognition and 0.06 for all the other processing steps which are independent on the number of objects. Assuming a linear growth of dictionaries, our system could recognize 40 objects with a mean computation time of 0.2s. In order to learn a much larger set of objects, the dictionary growth should be limited by introducing additional filtering of dictionaries in order to keep only the most frequently repeated features. Another approach could be to learn a fixed dictionary of visual features in a first phase, before learning the objects. Such approach would not be incremental as ours, but would make it possible to use much more efficient data structures as used in image retrieval (e.g., [34]) that would scale to a much larger number of objects.

The object representation and learning approach presented in this paper takes advantage of social interactions as these interactions produce object motions that are important in our system, but does not explicitly engage in such interactions. In a related work however, our system has been inte-

grated within a curiosity-based active object exploration architecture [32,51] that took advantage of the social environment by asking the human partner to manipulate a particular object. This was possible because our approach provides an assessment of an object model quality through its number of views and its recognition probability. This quality measure has been used to guide the choice of an object, an action, and an actor (i.e., the robot itself or the human partner) in order to explore based on the achieved learning progress.

This work made a number of engineering choices whose consequences can be questioned. Among these, the choice of a fixed external RGB-D sensor made it possible to simplify implementation, improve the quality of the data, and therefore the system performance. In particular, it avoids the complex problem of learning gaze control that involves eyes and neck joints that have not been considered in this work [38]. However, this removes the possibility for the system to control its gaze direction. Imagining the implementation of our system with a gaze-controlled camera on the robot head, our image processing stream should not be strongly affected (beside the loss of performance as illustrated in Section 5.2) as long as object tracking remains possible. The entities classification however will require improvements as it currently depends on the fact that the camera is static to analyse entities motions. A new component computing entities motion in the robot body frame would therefore be required. As an alternative for entity categorisation, we could extend our algorithm by including the head pose (the states of neck joints) and the gaze direction into the arm-torso dictionary. This modification will allow to consider the relation between the entity localization over time relative to the camera pose, thus allowing the camera motion. The calibration of the sensor currently performed by an initial calibration procedure could also be performed in a more natural way, following for example approaches learning visuo-motor coordination (e.g. [8,7]).

Concerning gaze control on the actual robot as a social cue, our engineering solution indeed makes it possible to make the robot look at an object or at humans for social interaction (thanks to the position of the object in the robot reference frame given by the RGB-D camera). However, the fact that the robot point of view from the external camera is not the point of view from the robot eyes which is assumed by humans can cause problems in human-robot interactions scenarios. Indeed, the human could assume that the side of the object seen by the robot is different from the one actually observed by the overhead camera.

Several parts of the proposed approach could also be extended by the use of more general learning approaches than the current hand-designed algorithms. For example, an interesting future work could be to replace the algorithm for entities categorisation proposed in Fig.8 by a more adaptive approach. A first step would be to learn the thresholds used in this procedure from data, but a more generic approach learning the entity behaviours and performing unsupervised categorisation of these behaviours to define the entities categories would be more appealing.

7 Conclusion and future work

We have developed a perceptual approach that enables a humanoid robot to explore its close environment in an interactive scenario, following the context of developmental learning. Without the use of image databases, pre-specified objects, known robot appearance or direct supervision but rather taking inspiration from infants development, the robot first learns by observing its surroundings, and then using its own interactive actions thanks to the identification of its own body.

This was achieved thanks to the integration of a generic physical entity appearance representation, a self- and others-identification capability, and actions for active exploration of the objects. The main lessons learned from this system are that:

- it is possible to make efficient models of all physical entities in front of a robot with a unified appearance model that can represent both textured objects such as the robot hands or soda cans and textureless objects such as toys or human hands,
- it is possible to categorize objects, human parts, and parts of the robot without prior knowledge on their appearances and using only their motion behaviour and its correlation with the robot proprioceptive sensing,
- the knowledge of these three categories are sufficient to update object models during manipulation, even when the object is in the robot hand, without the need of a precise body schema, nor initial knowledge of the robot appearance.

An interesting extension of this work would be to improve the integration of experience gathered by the robot through interaction with its environment into the processing pipeline itself. In infants, the development of capabilities to manipulate objects has an influence on their perception and especially attention [50]. It would be advantageous to implement a similar feature: once the robot has explored an object manually at a close scale, it has acquired more knowledge about the importance of its visual features for interaction or correct recognition. This experience could provide a feedback to the perceptual system, for example by changing the attention model or notion of saliency to be able to detect these objects at a greater distance.

Our developmental approach could be further extended by learning action primitives instead of using hand designed actions. While we focus on perception in this work, infants

develop simultaneously their recognition and action capabilities. It would be interesting to work on a more complete developmental approach for robots by learning the appropriate actions to manipulate the objects (following for example [38]) at the same time as learning to recognize these objects or as learning the affordances that make it possible to decide which actions apply to a given object. Learning these actions should be coupled with learning a more complete body schema than the simple partial body image that is learned in our current approach. Learning the full body schema would make it possible to extend self-recognition to more complex parts of the body of the robot, and would make it possible to perform more efficient manipulation actions.

Finally, it would also be interesting to extend our approach by integrating the audio information in our system. While seeking the multimodality of learning and taking inspiration from infant-directed interaction, when an adult names an object while showing it to the infant, we could learn about objects not only from visual data but also from audio information. This can be viewed as a step towards the development of common language between the robot and its human partner, where the robot is able to learn objects associated with any names that its user would like to use, while it could help to improve object recognition in more complex interactive scenarios.

Acknowledgments

This work was supported by the French ANR program (ANR-10-BLAN-0216) through Project MACSi, and partly by the European Commission, within the CoDyCo project (FP7-ICT-2011-9, No. 600716). The authors would like to thank the anonymous reviewers for their comments that greatly helped improving the quality of the paper.

References

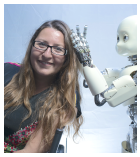
1. Aldavert, D., Ramisa, A., López de Mántaras, R., Toledo, R., et al.: Real-time object segmentation using a bag of features approach. Artificial Intelligence Research and Development pp. 321–329 (2010)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding **110**, 346–359 (2008)
3. Belongie, S., Carson, C., Greenspan, H., Malik, J.: Color-and texture-based image segmentation using EM and its application to content-based image retrieval. In: IEEE Conf. on Computer Vision, pp. 675–682. IEEE (1998)
4. Beucher, S., Meyer, F.: The morphological approach to segmentation: the watershed transformation. mathematical morphology in image processing. Optical Engineering **34**, 433–481 (1993)
5. Browatzki, B., Tikhanoff, V., Metta, G., Bulthoff, H., Wallraven, C.: Active object recognition on a humanoid robot. In: IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 2021–2028 (2012)
6. Burger, W., Burge, M.J.: Digital image processing. Springer (2008)
7. Chao, F., Lee, M.H., Jiang, M., Zhou, C.: An infant development-inspired approach to robot hand-eye coordination. Int. Journal of Advanced Robotic Systems **11**, 15 (2014)
8. Chinellato, E., Antonelli, M., Grzyb, B., del Pobil, A.: Implicit sensorimotor mapping of the peripersonal space by gazing and reaching. IEEE Trans. on Autonomous Mental Development **3**(1), 43–53 (2011)
9. Chu, V., McMahon, I., Riano, L., McDonald, C., He, Q., Martinez Perez-Tejada, J., Arrigo, M., Fitter, N., Nappo, J., Darrell, T., Kuchenbecker, K.: Using robotic exploratory procedures to learn the meaning of haptic adjectives. In: IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 3048–3055 (2013)
10. Crandall, D.J., Felzenszwalb, P.F., Huttenlocher, D.P.: Spatial priors for part-based recognition using statistical models. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10–17 (2005)
11. D. Schiebener J. Morimoto, T.A., Ude, A.: Integrating visual perception and manipulation for autonomous learning of object representations. Adaptive Behavior **21**(5), 328–345 (2013)
12. Dickscheid, T., Schindler, F., Förstner, W.: Coding images with local features. Int. Journal on Computer Vision **94**, 154–174 (2011)
13. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2Nd Edition). Wiley-Interscience (2000)
14. Everingham, M., Eslami, S., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. Journal of Computer Vision pp. 1–39 (2014)
15. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 264–271 (2003)
16. Fergus, R., Perona, P., Zisserman, A.: A sparse object category model for efficient learning and exhaustive recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 380–387 (2005)
17. Fiala, M.: Artag, a fiducial marker system using digital techniques. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 590–596 (2005)
18. Filliat, D.: A visual bag of words method for interactive qualitative localization and mapping. In: IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 3921–3926 (2007)
19. Gaël, G., Benoît, J.: Eigen v3. http://eigen.tuxfamily.org (2010)
20. Gevers, T., Smeulders, A.W.: Color-based object recognition. Pattern recognition **32**(3), 453–464 (1999)
21. Gold, K., Scassellati, B.: Learning acceptable windows of contingency. Connection Science **18**(2), 217–228 (2006)
22. Goldstein, E.B.: Sensation and perception. Wadsworth Publishing Company (2010)
23. Grauman, K., Leibe, B.: Visual Object Recognition. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers (2011)
24. Griffith, S., Sukhoy, V., Stoytchev, A.: Using sequences of movement dependency graphs to form object categories. In: IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids), pp. 715–720 (2011)
25. Gupta, M., Sukhatme, G.: Using manipulation primitives for brick sorting in clutter. In: IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 3883–3889 (2012)
26. Harman, K.L., Humphrey, G., Goodale, M.A.: Active manual control of object views facilitates visual recognition. Current Biology **9**(22), 1315 – 1318 (1999)
27. Hoffmann, M., Marques, H., Hernandez Arieta, A., Sumioka, H., Lungarella, M., Pfeifer, R.: Body schema in robotics: A review. IEEE Trans. on Autonomous Mental Development **2**(4), 304–324 (2010)
28. van Hoof, H., Kroemer, O., Peters, J.: Probabilistic segmentation and targeted exploration of objects in cluttered environments. IEEE Trans. on Robotics **30**(5), 1198–1209 (2014)

29. Huang, T., Yang, G., Tang, G.: A fast two-dimensional median filtering algorithm. IEEE Trans. on Acoustics, Speech and Signal Processing **27**(1), 13–18 (1979)
30. Hulse, M., McBrid, S., Lee, M.: Robotic hand-eye coordination without global reference: A biologically inspired learning scheme. In: IEEE Int. Conf. on Development and Learning (ICDL), pp. 1–6. IEEE (2009)
31. Ivaldi, S., Lyubova, N., Gérardeaux-Viret, D., Droniou, A., Anzalone, S.M., Chetouani, M., Filliat, D., Sigaud, O.: Perception and human interaction for developmental learning of objects and affordances. In: IEEE Int. Conf. on Humanoid Robots (Humanoids), pp. 248 – 254 (2012)
32. Ivaldi, S., Nguyen, S., Lyubova, N., Droniou, A., Padois, V., Filliat, D., Oudeyer, P.Y., Sigaud, O.: Object learning through active exploration. IEEE Trans. on Autonomous Mental Development (2013)
33. Ivaldi, S., Nguyen, S., Lyubova, N., Droniou, A., Padois, V., Filliat, D., Oudeyer, P.Y., Sigaud, O.: Object learning through active exploration. IEEE Transactions on Autonomous Mental Development **6**, 56–72 (2014)
34. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. International Journal of Computer Vision **87**(3), 316–336 (2010)
35. Kemp, C., Edsinger, A.: What can i control?: The development of visual categories for a robots body and the world that it influences. In: Int. Workshop on Epigenetic Robotics (Epirob), pp. 33–40 (2006)
36. Kraft, D., Pugeault, N., Baseski, E., Popovic, M., Kragic, D., Kalkan, S., Worgotter, F., Kruger, N.: Birth of the object: detection of objectness and extraction of object shape through object-action complexes. Int. Journal of Humanoid Robotics **05**(02), 247–265 (2008)
37. Krainin, M., Henry, P., Ren, X., Fox, D.: Manipulator and object tracking for in-hand 3D object modeling. Int. Journal of Robotics Research **30**(11), 1311–1327 (2011)
38. Law, J., Shaw, P., Lee, M., Sheldon, M.: From saccades to grasping: A model of coordinated reaching through simulated development on a humanoid robot. Autonomous Mental Development, IEEE Transactions on **6**(2), 93–109 (2014)
39. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 97 – 104 (2004)
40. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Int. Joint Conf. on Artificial Intelligence (IJCAI), pp. 674–679 (1981)
41. Lyubova, N.: Developmental approach of perception for a humanoid robot. Ph.D. thesis, ENSTA ParisTech (2013)
42. Marjanovic, M.J., Scassellati, B., Williamson, M.M.: Self-taught visually-guided pointing for a humanoid robot. In: From Animals to Animats 4: Int. Conf. on Simulation of Adaptive Behavior (SAB), pp. 35 – 44 (1996)
43. Metta, G., Fitzpatrick, P.M.: Better vision through manipulation. Adaptive Behaviour **11**(2), 109–128 (2003)
44. Michel, P., Gold, K., Scassellati, B.: Motion-based robotic self-recognition. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), vol. 3, pp. 2763–2768. IEEE (2004)
45. Micusik, B., Kosecka, J.: Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry. In: IEEE Int. Conf. on Computer Vision (ICCV), pp. 625–632 (2009)
46. Modayil, J., Kuipers, B.: The initial development of object knowledge by a learning robot. Robotics and Autonomous Systems **56**, 879–890 (2008)
47. Nagi, J., Ducatelle, F., Di Caro, G.A., Ciresan, D., Meier, U., Giusti, A., Nagi, F., Schmidhuber, J., Gambardella, L.M.: Max-pooling convolutional neural networks for vision-based hand gesture recognition. In: IEEE Int. Conf. on Signal and Image Processing Applications (ICSIPA), pp. 342–347 (2011)
48. Natale, L., Nori, F., Metta, G., Fumagalli, M., Ivaldi, S., Pattacini, U., Randazzo, M., Schmitz, A., Sandini, G.: The icub platform: a tool for studying intrinsically motivated learning. In: Intrinsically motivated learning in natural and artificial systems, pp. 433–458. Springer (2013)
49. Natale, L., Orabona, F., Berton, F., Metta, G., Sandini, G.: From sensorimotor development to object perception. In: IEEE/RAS Int. Conf. on Humanoid Robots, pp. 226–231 (2005)
50. Needham, A., Barrett, T., Peterman, K.: A pick-me-up for infants exploratory skills: Early simulated experiences reaching for objects using sticky mittens enhances young infants object exploration skills. Infant Behavior and Development **25**(3), 279 – 295 (2002)
51. Nguyen, S.M., Ivaldi, S., Lyubova, N., Droniou, A., Gérardeaux-Viret, D., Filliat, D., Padois, V., Sigaud, O., Oudeyer, P.Y.: Learning to recognize objects through curiosity-driven manipulation with the icub humanoid robot. In: Int. Conf. on Development and Learning, pp. 1 – 8 (2013)
52. Orabona, F., Metta, G., Sandini, G.: A proto-object based visual attention model. In: L. Paletta, E. Rome (eds.) Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint, *Lecture Notes in Computer Science*, vol. 4840, pp. 198–215. Springer Berlin Heidelberg (2007)
53. Piaget, J.: Play, dreams and imitation in childhood. Routledge, London (1999)
54. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3282–3289 (2012)
55. Pylyshyn, Z.W.: Visual indexes, preconceptual objects, and situated vision. Cognition **80**, 127–158 (2001)
56. Rensink, R.A.: Seeing, sensing, and scrutinizing. Vision research **40**(10-12), 1469–1487 (2000)
57. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1605–1614 (2006)
58. Saegusa, R., Metta, G., Sandini, G.: Body definition based on visuomotor correlation. IEEE Trans. on Industrial Electronics **59**(8), 3199–3210 (2012)
59. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 593 – 600 (1994)
60. Shih, F.Y.: Image processing and mathematical morphology: Fundamentals and applications. CRC PressI Llc (2009)
61. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE (2008)
62. Sinapov, J., Bergquist, T., Schenck, C., Ohiri, U., Griffith, S., Stoytchev, A.: Interactive object recognition using proprioceptive and auditory feedback. I. J. Robotic Res. **30**(10), 1250–1262 (2011)
63. Sivic, J., Zisserman, A.: Video google: Text retrieval approach to object matching in videos. In: Int. Conf. on Computer Vision, vol. 2, pp. 1470–1477 (2003)
64. Southey, T., Little, J.J.: Object discovery through motion, appearance and shape. In: AAAI Workshop on Cognitive Robotics, p. 9 (2006)
65. Spelke, E.S.: Principles of object perception. Cognitive Science **14**, 29–56 (1990)
66. Spelke, E.S., Kinzler, K.D.: Core knowledge. Developmental Science **10**(1), 89–96 (2007)
67. Torres-Jara, E., Natale, L., Fitzpatrick, P.: Tapping into touch. In: Int. Workshop on Epigenetic Robotics (Epirob), pp. 79–86. Lund University Cognitive Studies (2005)

68. Ude, A., Omrčen, D., Cheng, G.: Making object learning and recognition an active process. Int. Journal of Humanoid Robotics **5**(02), 267–286 (2008)
69. Viola, P., Jones, M.J.: Robust real-time face detection. Int. Journal on Computer Vision **57**, 137–154 (2004)
70. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19**(9), 1395–407 (2006)
71. Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., Thelen, E.: Autonomous mental development by robots and animals. Science **291**(5504), 599–600 (2001)
72. Wersing, H., Kirstein, S., Götting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J.J., Ritter, H., Körner, E.: Online learning of objects in a biologically motivated visual architecture. Int. J. Neural Systems **17**(4), 219–230 (2007)
73. Yang, M.H., Ahuja, N.: Gaussian mixture model for human skin color and its application in image and video databases. In: SPIE: Storage and Retrieval for Image and Video Databases, vol. 3656, pp. 458–466 (1999)
74. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–10 (2012)

Natalia Lyubova obtained a PhD in Cognitive Robotics from Ecole Nationale Suprieure de Techniques Avances ParisTech and Ecole Polytechnique in 2013. She received her Engineering degree from the Northern Arctic Federal University (Russia) in 2008, and M.S. degree from the University of Eastern Finland (Finland) in 2010. Since March 2014, she is a researcher at Aldebaran-Robotics, Perception team. Her main research interest is learning through perception and action on a robot.
`http://www.ensta-paristech.fr/~lyubova`

David Filliat graduated from the Ecole Polytechnique in 1997 and obtained a PhD on bio-inspired robotics navigation from Paris VI university in 2001. After 4 years as an expert for the robotic programs in the French armament procurement agency, he is now professor at Ecole Nationale Supèrieure de Techniques Avancées ParisTech. Head of the Robotics and Computer Vision team, he obtained the Habilitation à Diriger des Recherches en 2011. His main research interest are perception, navigation and learning in the frame of the developmental approach for humanoid and mobile robotics.
`http://www.ensta-paristech.fr/~filliat`



Serena Ivaldi Serena Ivaldi is a researcher in INRIA. She received the B.S. and M.S. degree in Computer Engineering, both with highest honors, at the University of Genoa (Italy) and her PhD in Humanoid Technologies in 2011, jointly at the University of Genoa and the Italian Institute of Technology. There she also held a research fellowship in the Robotics, Brain and Cognitive Sciences Department. She was a postdoctoral researcher in the Institut des Systèmes Intelligents et de Robotique (ISIR) in University Pierre et Marie Curie, Paris, then in the Intelligent Autonomous Systems Laboratory in the Technical University of Darmstadt, Germany. Since November 2014, she is a researcher in INRIA Nancy Grand-Est. Her research is centered on robots interacting physically and socially with humans, blending learning, perception and control.
`http://www.loria.fr/~sivaldi/`